Calhoun: The NPS Institutional Archive

DSpace Repository

Faculty and Researchers | Faculty and Researchers' Publications

2020-07

Filtered Poisson Process Bandit on a Continuum

Grant, James A.; Szechtman, Roberto

ArXiv

Grant, James A., and Roberto Szechtman. "Filtered Poisson Process Bandit on a Continuum." arXiv preprint arXiv:2007.09966 (2020).
http://hdl.handle.net/10945/65342

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun

http://www.nps.edu/library

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

Filtered Poisson Process Bandit on a Continuum

James A. Grant*[1] and Roberto Szechtman†[2]

[1]Department of Mathematics and Statistics, Lancaster University, UK
[2]Department of Operations Research, Naval Postgraduate School, CA, USA

July 20, 2020

Abstract

We consider a version of the continuum armed bandit where an action induces a filtered realisation of a non-homogeneous Poisson process. Point data in the filtered sample are then revealed to the decision-maker, whose reward is the total number of revealed points. Using knowledge of the function governing the filtering, but without knowledge of the Poisson intensity function, the decision-maker seeks to maximise the expected number of revealed points over T rounds. We propose an upper confidence bound algorithm for this problem utilising data-adaptive discretisation of the action space. This approach enjoys $\tilde{O}(T^{2/3})$ regret under a Lipschitz assumption on the reward function. We provide lower bounds on the regret of any algorithm for the problem, via new lower bounds for related finite-armed bandits, and show that the orders of the upper and lower bounds match up to a logarithmic factor.
Keywords: Applied Probability; Poisson Processes; Multi-Armed Bandit; Machine Learning

1 Introduction

The challenge of detecting interesting events, using limited resources, arises in numerous settings. In a defence context, surveillance teams wish to observe suspicious activity or gain intelligence. In ecological and environmental data collection, scientists wish to observe behaviours of endangered species or record notable measurements of environmental variables. In manufacturing and logistics settings, it is desirable to observe faults in machine operation or a supply chain.

However, in all of these settings, practitioners may face the problem of having insufficient resource to observe everything they wish to, and must optimise their resource allocation to maximise the detection of events. In these settings "resource" may refer to human searchers, fixed or mobile sensors, cameras, or a variety of other equipment with a capacity to observe events of interest.

*j.grant@lancaster.ac.uk; corresponding author
†rszechtm@nps.edu

Two factors play a particularly important role in the rate of detection. Crudely put, these are where we look, and how good we are at looking. In any of these settings, we can only expect to observe events in locations (spatial or temporal) where we deploy resource. Further, the precision of the detection may also be affected by how resource is deployed. If resource is spread over a large region, the probability of detecting events within this region may be lower than if focused on a small area.

Inspired by these challenges, we consider a stylised model of resource allocation which captures the challenge of balancing coverage and detection probability. This framework is sufficiently abstract to model problems across the various aforementioned applications and beyond.

Consider a decision-maker who aims to detect the maximum number of events occurring according to a Non-homogeneous Poisson process (NHPP) on a segment $[0, 1]$. The decision-maker selects a point $y \in [0, 1]$ and then sweeps the sub-segment $[0, y]$ searching for events. However, the decision-maker's search is imperfect, in that events in $[0, y]$ are detected, independently of each other, with *filtering probability* $\gamma(y)$, where $\gamma : [0, 1] \to [0, 1]$, is a known, nonincreasing function. The expected number of events detected by the decision-maker on a single sweep is then determined by the filtering probability, and the cumulative intensity function (CIF) of the NHPP,

$$\Lambda(y) = \int_0^y \lambda(z)dz, \ \ \forall\ y \in [0, 1]$$

where $\lambda : [0, 1] \to \mathbb{R}$ is the rate function of the NHPP. Given the decision-maker chooses to sweep $[0, y]$, the number of events detected has a Poisson$(\Lambda(y)\gamma(y))$ distribution.

Figure 1 illustrates this process. An example intensity function λ is represented by the blue curve and a function γ giving the filtering probability is given by the black curve. The blue points towards the bottom of the left pane illustrate a single sample of events from the NHPP with intensity λ. The decision-maker selects $y = 0.6$ and sweeps the sub-segment $[0, 0.6]$, detecting each event therein with probability $\gamma(0.6)$. The red piecewise-constant function in the right pane illustrates the effective filtering probability over $[0, 1]$. The points plotted in red then represent the events actually detected by the decision-maker during their imperfect search - which we observe are a subset of the events that actually arose.

In this paper, we consider a sequential variant of this problem, where the CIF, Λ, is unknown to the decision-maker, but the choice of endpoint y can be updated over a series of rounds, in response to observing the locations of detected events in previous rounds. The decision-maker's aim is then to maximise the expected number of detected events over $T \in \mathbb{N}$ rounds. The study of this problem is motivated both by its theoretical challenge and its practical interest.

Versions of this problem may arise in a number of settings such as ecological surveillance, defence, and logistics, where sightings of endangered species, criminal activity, or machine faults may for instance comprise the events of interest. As a motivating, and sufficiently general example, consider a scenario where observations are made by searchers (representing cameras, sensors, robotic and human searchers, etc.), that must restart at the same point after each round. We note that while in the material that follows we will treat the line segment as indexing space (for clarity and consistency), it could equivalently be thought of as indexing time or space-time and apply to a yet broader range of examples.

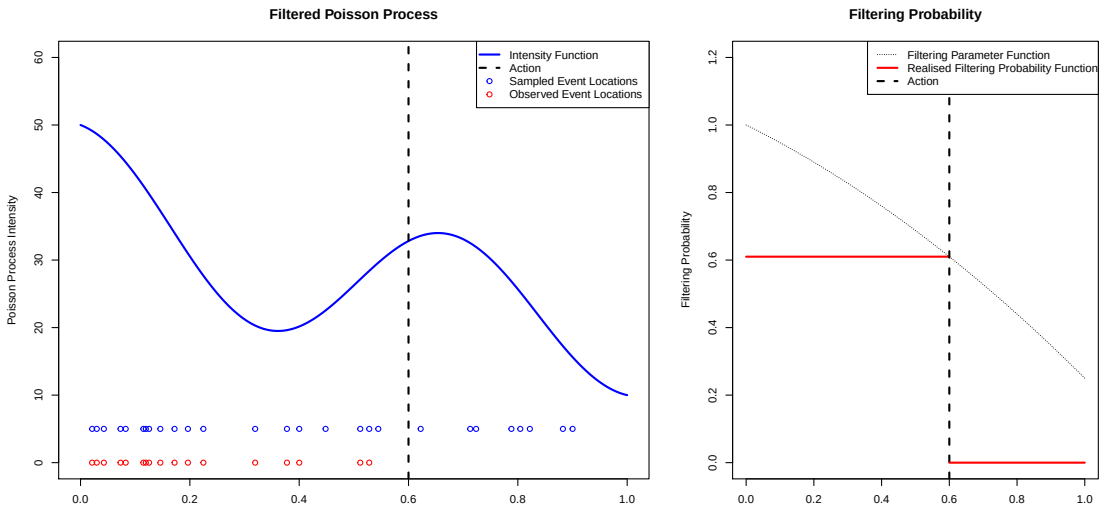

Figure 1: Graphical representation of the filtering process.

From a theoretical perspective, the problem is closely related to the one-dimensional case of the stochastic continuum-armed bandit (CAB) problem (Agrawal 1995). This is a sequential decision-making problem where in each of a series of rounds $t \in [T] \equiv \{1, \ldots, T\}$, a decision-maker selects an action $x_t \in [0, 1]$ and receives a reward, which is a noisy realisation of some unknown smooth function $f : [0, 1] \to [0, 1]$ evaluated at x_t. The decision-maker's aim is to maximise the expected sum of rewards amassed over T rounds. To realise this aim, the decision-maker must deploy a strategy which appropriately balances between exploring the action space $[0, 1]$ to learn the function f, and exploiting this information, selecting actions known to produce larger rewards to maximise the cumulative total.

In the Poisson process-based problem at hand, a similar dilemma arises, we lack knowledge of the filtered CIF - which corresponds to the reward function - and can only hope to maximise the sum of rewards by exploring the action space - i.e. choosing a range of endpoints $y \in [0, 1]$. However, the feedback received on actions in our problem is much richer than in the standard CAB problem. In addition to a noisy realisation of the filtered CIF, $\Lambda\gamma$, we observe the location of detected events, which can help with the estimation of the reward function beyond the inferences from smoothness properties alone. Methods for the standard CAB problem are therefore inappropriate for the problem we face, as is the existing unmodified theory. In this paper we present a specific treatment of the previously described sequential endpoint selection problem, which we henceforth refer to as a Filtered Poisson Process Bandit (FPPB), deriving a bespoke decision-making algorithm and theoretical analysis of the problem.

1.1 Related Literature

Sequential decision-making problems on continuous action spaces have been studied extensively, following from initial works of Agrawal (1995) and Kleinberg (2005). Most successful strategies

have employed a combination of adaptive discretisation of the action space, and optimism in the face of uncertainty. Our approach for the FPPB problem, also uses these techniques.

Adaptive discretisation, as used in the "Zooming" algorithm of Kleinberg et al. (2008) and "hierarchical online optimisation" (HOO) algorithm of Bubeck et al. (2011a), reduces the available action space in round t to some $\mathcal{A}_t \subset [0, 1]$. Restricting the action set ensures exploration occurs at a predictable rate, and makes the action selection more straightforward. Gradually, as the rounds proceed and more information is gathered, $\mathcal{A}_t$ is increased, usually in a data-adaptive fashion to permit choice from a more granular set of actions. Intuitively, this is also appealing, as when estimates of the reward are very crude, there is little motivation to make decisions at a very granular level.

Optimistic approaches are those which encourage an appropriate balance of exploration and exploitation by making decisions with respect to high probability upper confidence bounds (UCBs) on the expected reward of the available actions. The Zooming and HOO algorithms both calculate UCBs for the reward of available actions in each round and select the action with the largest UCB. These approaches were the first to achieve order optimal performance, in terms of regret, for this class of problems.

Strong results have also been obtained by approaches which use Gaussian processes and avoid discretisation of the action space. The GP-UCB (Gaussian Process - Upper Confidence Bound) algorithm of Srinivas et al. (2010) constructs an upper confidence bound on the reward function over all actions, rather than at specific points, and selects the action which maximises this UCB function. This method also has order optimal performance guarantees, but with respect to a Bayesian measure of regret, rather than the frequentist one used in the analysis of the Zooming and HOO algorithms.

It is worth noting that none of these algorithms can sensibly be applied to the FPPB, and that their theoretical guarantees do not carry to the FPPB problem. Principally, this is because they lack a means to handle the additional feedback in terms of the location data, but a more subtle point is that without modification, these methods are not suited to unbounded rewards, as we have in this setting, with the Poisson distributed reward.

Grant et al. (2020) consider a filtered Poisson bandit problem which is similar in some senses to ours, but theirs employs a fixed discretisation of the action space, such that the spatial locations of the events are irrelevant. They focus instead on the challenges of choosing multiple non-overlapping sub-segments and analyse performance with respect to the best possible action among a fixed discrete set. Grant et al. (2019) considers a continuous action space, but without filtering of the observations. Inference is therefore more straightforward in this setting, and the Thompson Sampling method proposed is not applicable to the FPPB setting. Recently, Lu et al. (2019) provide an algorithm combining the adaptive discretisation of Kleinberg et al. (2008) and heavy tailed UCBs of Bubeck et al. (2013) for a version of the CAB problem with heavy-tailed reward noise distributions. While the Poisson does fit in to this class of distributions, it also enjoys tighter bespoke concentration results, and a general heavy-tailed approach is overly conservative for the FPPB - even if event locations were not observed.

1.2 Key Contributions and Structure

The main contribution is a UCB algorithm with $\tilde{O}(T^{2/3})$ regret over T rounds. By derivation of a lower bound, we show that under the assumptions on the CIF, this is optimal up to a logarithmic factor. From the methodological viewpoint, we extend the Lipschitz multi-armed bandit framework (Kleinberg et al. 2008) to deal with a filtered Poisson process on continuum.

The remainder of the paper is structured as follows. In Section 2 we precisely state the problem of interest. In Section 3 we present our UCB approach to the problem. Sections 4 and 5 provide the upper and lower bounds on regret respectively. We conclude with a simulation of our method in Section 6, and discussion in Section 7.

2 Model

The formal specification of the FPPB problem is as follows. In rounds $t \in [T]$, the decision-maker selects an endpoint $y_t \in [0,1]$ and makes an observation on the sub-segment $[0, y_t]$. The environment generates a realisation of the NHPP with CIF Λ, consisting of an increasing sequence of event locations $\{X_{t,1}, X_{t,2}, \ldots, X_{t,N_t}\} \in [0,1]^{N_t}$, where $N_t \sim \text{Poisson}(\Lambda(1))$. The end-point selected by the decision-maker implies a filtering probability $\gamma(y_t) \in [0,1]$, such that events to the left of y_t are detected independently of each other with probability $\gamma(y_t)$, and all events to the right of y_t are not detected. As a result, a sequence of i.i.d. Bernoulli$(\gamma(y_t))$ random variables, $B_1, B_2, \ldots, B_{N_t}$, is generated. The decision maker receives the count of detected events $R_t \equiv R_t(y_t) = \sum_{k=1}^{N_t} \mathbb{1}(B_{t,k} = 1, X_{t,k} \leq y_t)$ as a reward, and observes the locations of detected events $X_{t,k}$ with $B_{t,k} = 1$ and $X_{t,k} \leq y_t$. By construction, $R_t \sim \text{Poisson}(\Lambda(y_t)\gamma(y_t))$.

The decision-maker's objective is to maximise the sum of rewards obtained over T rounds, $\sum_{t=1}^{T} R_t$. To realise this objective we aim to determine a *policy*, A, which maps from a history of actions and observations to a next action, which maximises the expected reward, or equivalently minimises the *regret*,

$$Reg_A(T) = \mathbb{E}\left(\sum_{t=1}^{T} R_t(z^*) - R_t(y_t)\right), \tag{1}$$

where $z^* \in \text{argmax}_{y\in[0,1]} \Lambda(y)\gamma(y)$ is an optimal endpoint which maximises the expected per-round reward. Here the expectation is with respect to both the random process governing the generation and filtering of events and the decision-maker's actions. We will be interested in upper bounding the regret as a function of T for our proposed algorithm, and comparing the order of this upper bound to that of lower bounds on the best achievable regret of any algorithm.

Bounded regret is achievable only if the reward function is suitably well-behaved as to admit learning from a finite sample of observations. This is ensured through assumptions on the form of the CIF and filtering function. These assumptions, enforced throughout the paper, are Lipschitz continuity of the filtered CIF and a rate bound,

A1: $|\gamma(y)\Lambda(y) - \gamma(x)\Lambda(x)| \leq m|y - x|, \forall x, y \in [0,1]$,

A2: $\lambda(y) \leq \lambda_{\max}$,

for $m, \lambda_{\max} \geq 0$ known and finite. Assumptions A1–A2 are used to bound the estimation error for the expected number of detected events in each cell; this can be achieved by including in the cell index an additive term proportional to the cell length. We also assume that $\gamma_{\min} = \inf_{y\in(0,1]}\{\gamma(y) > 0\} > 0$; this is without loss of generality, as segments with $\gamma(\cdot) = 0$ do not contain the optimal endpoint.

3 Algorithm

In this section we present our algorithm for the FPPB problem, CIF-UCB, given as Algorithm 1.

Algorithm 1 CIF-UCB (Cumulative Intensity Function - Upper Confidence Bound)

Input: Rate bound $\lambda_{\max}$, filtering probabilities $\gamma(\cdot)$, Lipschitz constant m, active cell set $\mathcal{A}_1 = \{(0, 1]\}$, effective number of samples $V_1(0, 1) = \emptyset$, index $\mathcal{I}_1(0, 1) = m$.
for $t = 1$ **to** T **do**
  Selection Rule:
  Find cell

$$(a_t, b_t] = \operatorname*{argmax}_{(x,y]\in\mathcal{A}_t} \mathcal{I}_t(x, y),$$

  breaking ties randomly.
  Do a sweep up to b_t.
  Update $V_{t+1}(a_t, b_t) = V_t(a_t, b_t) \cup \{t\}$, and

$$\zeta_{t+1}(b_t) = \frac{6\max\{1, \lambda_{\max}\}\log(T)}{\sum_{i=1}^{|V_{t+1}(a_t,b_t)|}\gamma(b_{\tau_i})} + \sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V_{t+1}(a_t,b_t)|}\gamma(b_{\tau_i})}}.$$

  Update $\bar{\Lambda}_{t+1}(b_t)$ as in (2).
  Division Rule:
  if $m(b_t - a_t) \geq \zeta_{t+1}(b_t)$ **then**
    Update the active cell set $\mathcal{A}_{t+1} = \mathcal{A}_t \setminus \{(a_t, b_t]\} \cup \{(a_t, (a_t + b_t)/2], ((a_t + b_t)/2, b_t]\}$.
    Set $V_{t+1}((a_t, (a_t + b_t)/2) = V_{t+1}(((a_t + b_t)/2, b_t) = V_{t+1}(a_t, b_t)$, and

$$\zeta_{t+1}\left(\frac{a_t + b_t}{2}\right) = \frac{6\max\{1, \lambda_{\max}\}\log(T)}{\sum_{i=1}^{|V_{t+1}(a_t,(a_t+b_t)/2)|}\gamma(b_{\tau_i})} + \sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V_{t+1}(a_t,(a_t+b_t)/2)|}\gamma(b_{\tau_i})}},$$

$$\zeta_{t+1}(b_t) = \frac{6\max\{1, \lambda_{\max}\}\log(T)}{\sum_{i=1}^{|V_{t+1}((a_t+b_t)/2,b_t)|}\gamma(b_{\tau_i})} + \sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V_{t+1}((a_t+b_t)/2,b_t)|}\gamma(b_{\tau_i})}}.$$

    Define $\bar{\Lambda}_{t+1}((a_t + b_t)/2)$ and $\bar{\Lambda}_{t+1}(b_t)$ as in (2).
  end if
  UCB Computation:
  Set $\mathcal{I}_{t+1}(x, y) = \gamma(y)\bar{\Lambda}_{t+1}(y) + m(y - x) + \gamma(y)\zeta_{t+1}(y)$ for all cells $(x, y] \in \mathcal{A}_{t+1}$.
end for

At a high level, CIF-UCB proceeds as follows. For each round $t = 1, \ldots, T$, the algorithm

maintains a set of active cells, $\mathcal{A}_t$, which form a partition of $[0,1]$. An index, $\mathcal{I}_t$, taking the form of optimistic estimate of the expected reward, is computed for each cell in $\mathcal{A}_t$. The algorithm selects the right endpoint of the active cell with largest index as the action for that round. Initially, the active set contains the unit interval, $\mathcal{A}_1 = \{(0,1]\}$, so that the algorithm does a complete sweep in the first round. If the number of sweeps of a cell exceeds some threshold in relation to its length, the cell is split in half. Hence, active cells make up a partition of the interval $[0,1]$ for all rounds. A new cell inherits the number of sweeps and detection count that fall in its interval from the parent cell.

Accumulating rewards over the interval to the left of the selected endpoint makes the problem structure combinatorial in nature, which poses a challenge for the analysis. The insight that makes the analysis tractable is that, by the independent increment property of the Poisson process, the filtered Poisson counts corresponding to the active cells that lie to the left of the endpoint selected by the algorithm in each round are independent. This leads to a CIF estimator for each active cell with tight error bounds.

We complete the notation needed to define the CIF estimator. Let $\{\mathcal{F}_t\}_{t=1}^T$ be the filtration induced by the sequence of event locations and cell selections $((a_t, b_t])_{t=1}^T$. Also, let

$$V_t(x,y) = \{\tau_1, \tau_2, \ldots\} \subseteq [t]$$

be the collection of (random) times when active cell $(x,y]$ is swept by round t and let,

$$Z_{\tau_i}(y) = \sum_{k=1}^{N_{\tau_i}} \mathbb{1}(B_{\tau_i,k} = 1, X_{\tau_i,k} \leq y)$$

be the filtered Poisson count to the left of y in round τ_i. Finally, let $\sum_{i=1}^{|V_t(x,y)|} Z_{\tau_i}(y)$ be the total filtered Poisson count to the left of y over the rounds when cell $(x,y]$ is swept. When the context is clear, we write V in lieu of $V_t(x,y)$

For active cell $(x,y]$, $\Lambda(y)$ is estimated by dividing the cumulative filtered Poisson counts up to y by its effective number of sweeps by round t,

$$\bar{\Lambda}_t(y) = \frac{\sum_{i=1}^{|V_t(x,y)|} Z_{\tau_i}(y)}{\sum_{i=1}^{|V_t(x,y)|} \gamma(b_{\tau_i})}. \tag{2}$$

Essentially, in (2) the filtered Poisson count is *unfiltered* by dividing it by $\sum_{i=1}^{|V_t(x,y)|} \gamma(b_{\tau_i})$. It's easy to see that $\bar{\Lambda}_t(y)$ is an unbiased estimator of $\Lambda(y)$.

CIF-UCB samples from the origin to the endpoint of the active cell with largest index, and divides the latter cell if its length exceeds certain threshold. The complexity of the CIF-UCB is $O(T)$ for the variable updates, and $O(\sum_{t=1}^T t \log t) = O(T^2 \log T)$ for sorting the indices, since there are at most t active cells by round t.

4 Upper Bound on Regret

In this section we present the first of our main theoretical contributions, an upper bound on the regret of CIF-UCB.

Theorem 1. *The regret of CIF-UCB applied to the FPPB problem, with CIF and filtering function satisfying Assumptions A1 and A2 satisfies*

$$Reg(T) = \tilde{O}(T^{2/3}).$$

Proof. The proof has three main stages. We first bound the CIF estimator error for each active cell (Lemma 1), and then use the Lipschitz assumption to extend the bound to include all the points inside an active cell (knowing that one of these points is an optimal endpoint for some active cell; Corollary 1). Second, we use the Division rule to express the confidence bound of each active cell in terms of its length (Lemma 2), which yields a bound for the per-round regret of the cell selected by the algorithm (Lemma 3). Finally, we accumulate these per-round regrets to obtain an upper bound for the regret over T rounds.

Firstly, we present the following concentration result, which asserts that the difference between the true CIF and the estimated CIF is unlikely to exceed the upper confidence terms used in Algorithm 1.

Lemma 1. *Let $(x, y]$ be an active cell in round t. Then,*

$$P\left(\left|\bar{\Lambda}_t(y) - \Lambda(y)\right| > \zeta_t(y)\right) \leq 2T^{-2},$$

where

$$\zeta_t(y) = \frac{6\log(T)\max\{1, \lambda_{\max}\}}{\sum_{i=1}^{|V_t(x,y)|}\gamma(b_{\tau_i})} + \sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V_t(x,y)|}\gamma(b_{\tau_i})}}.$$

Proof. The Poisson count $Z_{\tau_i}(y)$ is $\mathcal{F}_{\tau_i}$ measurable and,

$$E[Z_{\tau_i}(y)|\mathcal{F}_{\tau_{i-1}}] = \Lambda(y)\gamma(b_{\tau_i}), \text{ a.s.}$$

Defining,

$$M_k(y) = \sum_{i=1}^{k}(Z_{\tau_i}(y) - \Lambda(y)\gamma(b_{\tau_i})),$$

it follows that $\{M_{k\wedge|V|}(y), \mathcal{F}_{\tau_k}\}_{k\geq 1}$ is a martingale, and $M_{k\wedge|V|}(y) - M_{(k-1)\wedge|V|}(y) = (Z_{\tau_k}(y) - \Lambda(y)\gamma(b_{\tau_i}))\mathbb{1}(k \leq |V|)$ is a martingale difference sequence. By Lemma 1 in (Grant et al. 2020),

$$P\left(\sum_{i=1}^{k\wedge|V|}(Z_{\tau_i}(y) - \Lambda(y)\gamma(b_{\tau_i})) > \eta\right) \leq \exp\left(-\frac{\eta^2}{2\Lambda(y)\sum_{i=1}^{|V|}\gamma(b_{\tau_i}) + 2\max\{1, \Lambda(y)\}\eta}\right).$$

Solving for the r.h.s. above equal to T^{-3} leads to,

$$\eta = 3\log(T)\max\{1,\Lambda(y)\} + \sqrt{(3\log(T)\max\{1,\Lambda(y)\})^2 + 6\Lambda(y)\log(T)\sum_{i=1}^{|V|}\gamma(b_{\tau_i})}$$
$$\leq 6\log(T)\max\{1,\lambda_{\max}\} + \sqrt{6\lambda_{\max}\log(T)\sum_{i=1}^{|V|}\gamma(b_{\tau_i})}.$$

It follows that the probability that

$$\sum_{i=1}^{k\wedge|V|}(Z_{\tau_i}(y) - \Lambda(y)\gamma(b_{\tau_i})) > 6\log(T)\max\{1,\lambda_{\max}\} + \sqrt{6\lambda_{\max}\log(T)\sum_{i=1}^{|V|}\gamma(b_{\tau_i})},$$

is at most T^{-3} for each $k \leq T$. Taking a union bound over all $k \leq T$, and replacing for the definition of $\bar{\Lambda}_t(y)$ and $\zeta_t(y)$ results in

$$P\left(\bar{\Lambda}_t(y) - \Lambda(y) > \zeta_t(y)\right) \leq T^{-2}.$$

Finally, using the same approach it can be shown that

$$P\left(\bar{\Lambda}_t(y) - \Lambda(y) < -\zeta_t(y)\right) \leq T^{-2},$$

so the proof is complete.

□

The Lipschitz assumption can be used to extend this to a high probability bound on the filtered CIF for active cells.

Corollary 1. *Let $(x,y] \in \mathcal{A}_t$. Then, with probability at least $1 - 2T^{-2}$*

$$\sup_{x<c\leq y}\left|\gamma(y)\bar{\Lambda}_t(y) - \gamma(c)\Lambda(c)\right| \leq m(y-x) + \gamma(y)\zeta_t(y).$$

Proof. By the Lipschitz assumption,

$$\sup_{x<c\leq y}|\gamma(y)\Lambda(y) - \gamma(c)\Lambda(c)| < m(y-x).$$

Hence,

$$P\Big(\sup_{x<c\leq y}\left|\gamma(y)\bar{\Lambda}_t(y) - \gamma(c)\Lambda(c)\right| > m(y-x) + \gamma(y)\zeta_t(y)\Big) \leq P(|\bar{\Lambda}_t(y) - \Lambda(y)| > \zeta_t(y)).$$

□

The index of a cell $(x,y]$ active in round t is

$$\mathcal{I}_t(x,y) = \gamma(y)\bar{\Lambda}_t(y) + m(y-x) + \gamma(y)\zeta_t(y).$$

The $\gamma(y)\bar{\Lambda}_t(y)$ part of the index induces exploitation, while the $m(y-x)+\gamma(y)\zeta_t(y)$ term promotes exploration.

All the results that follow in this section are on the sample paths where

$$\sup_{x<c\le y}\left|\gamma(y)\bar{\Lambda}_t(y)-\gamma(c)\Lambda(c)\right| \le m(y-x)+\gamma(y)\zeta_t(y) \tag{3}$$

holds for all rounds $t=1,\dots,T$. By Corollary 1, the contribution to the regret of the sample paths that violate (3) is of order $O(1)$, after accounting for the T rounds and up to T cells by round T.

Our next result bounds the upper confidence term ζ_t for an active cell on the high probability event of Corollary 1.

Lemma 2. *For $(x,y]\in\mathcal{A}_t$,*

$$\zeta_t(y) \le 4m^2(y-x)^2\max\{1,1/\lambda_{\max}\}+2m(y-x).$$

Proof. Let $V^{(p)}(x,y)$ be the set of rounds the parent cell of $(x,y]$ got swept. The Division rule for the parent cell implies

$$2m(y-x) \ge \frac{6\max\{1,\lambda_{\max}\}\log(T)}{\sum_{i=1}^{|V^{(p)}(x,y)|}\gamma(b_{\tau_i})}+\sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V^{(p)}(x,y)|}\gamma(b_{\tau_i})}}.$$

Then, we obtain the conservative lower bound,

$$\sum_{i=1}^{|V^{(p)}(x,y)|}\gamma(b_{\tau_i}) \ge \frac{3\lambda_{\max}\log(T)}{2m^2(y-x)^2}. \tag{4}$$

Next we upper bound $\zeta_t(y)$,

$$\begin{aligned}\zeta_t(y) &\le \frac{6\max\{1,\lambda_{\max}\}\log(T)}{\sum_{i=1}^{|V^{(p)}(x,y)|}\gamma(b_{\tau_i})}+\sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V^{(p)}(x,y)|}\gamma(b_{\tau_i})}}\\ &\le 4m^2(y-x)^2\max\{1,1/\lambda_{\max}\}+2m(y-x),\end{aligned}$$

where the first inequality follows from the definition of $\zeta_t(y)$, and the second inequality follows from the lower bound (4).

□

Let z^* be an optimal endpoint (i.e., $\gamma(z^*)\Lambda(z^*)\ge\gamma(y)\Lambda(y)$ for all $y\in[0,1]$), and $(u_t,v_t]\in\mathcal{A}_t$ the cell that contains z^* in round t. The next result bounds the regret

$$\Delta(a_t,b_t)=\gamma(z^*)\Lambda(z^*)-\gamma(b_t)\Lambda(b_t),$$

incurred in each round in terms of the length of the cell selected by the algorithm.

Lemma 3. *The round t regret $\Delta(a_t, b_t)$ satisfies*

$$\Delta(a_t, b_t) \leq 8m^2(b_t - a_t)^2 \max\{1, 1/\lambda_{\max}\} + 5m(b_t - a_t).$$

Proof. We will show that

$$\gamma(z^*)\Lambda(z^*) \leq \mathcal{I}_t(a_t, b_t) \leq \gamma(b_t)\Lambda(b_t) + 5m(b_t - a_t) + 8m^2(b_t - a_t)^2 \max\{1, 1/\lambda_{\max}\},$$

from where the claim follows.

For the first inequality, we observe that,

$$\begin{aligned}\mathcal{I}_t(a_t, b_t) \geq \mathcal{I}_t(u_t, v_t) &\geq \gamma(v_t)\Lambda(v_t) + m(v_t - u_t) \\ &\geq \gamma(v_t)\Lambda(v_t) + m(v_t - z^*) \geq \gamma(z^*)\Lambda(z^*).\end{aligned}$$

In order, these inequalities follow from the Selection rule, the definition of the index function $\mathcal{I}_t$ and Corollary 1, the fact that $z^* \in (u_t, v_t]$, and the Lipschitz assumption. In the other direction, we have by application of Corollary 1, and then Lemma 2,

$$\begin{aligned}\mathcal{I}_t(a_t, b_t) &\leq \gamma(b_t)\Lambda(b_t) + m(b_t - a_t) + 2\gamma(b_t)\zeta_t(b_t) \\ &\leq \gamma(b_t)\Lambda(b_t) + 5m(b_t - a_t) + 8m^2(b_t - a_t)^2 \max\{1, 1/\lambda_{\max}\}.\end{aligned}$$

□

The final stage of the proof combines these results to realise the bound on regret. By Lemma 3, the regret of cells with length at most ℓ is bounded by

$$T(8m^2\ell^2 \max\{1, 1/\lambda_{\max}\} + 5m\ell) \tag{5}$$

over all rounds.

Cells with final length ℓ have three properties: (i) there are at most $1/\ell$ such cells; (ii) their regret per round is at most $8m^2\ell^2 \max\{1, 1/\lambda_{\max}\} + 5m\ell$ (Lemma 3); and (iii), satisfy (Division rule)

$$m\ell \leq \frac{6\log(T)\max\{1, \lambda_{\max}\}}{\sum_{i=1}^{|V|}\gamma(b_{\tau_i})} + \sqrt{\frac{6\lambda_{\max}\log(T)}{\sum_{i=1}^{|V|}\gamma(b_{\tau_i})}}.$$

Solving the quadratic inequality leads, after some algebra, to

$$\sum_{i=1}^{|V|}\gamma(b_{\tau_i}) \leq \frac{12\max\{1, \lambda_{\max}\}\log(T)}{m\ell} + \frac{6\lambda_{\max}\log(T)}{m^2\ell^2}.$$

Since $|V|\gamma_{\min} \leq \sum_{i=1}^{|V|}\gamma(b_{\tau_i})$, the number of selections is bounded above by

$$|V| \leq \frac{12\max\{1, \lambda_{\max}\}\log(T)}{\gamma_{\min} m\ell} + \frac{6\lambda_{\max}\log(T)}{\gamma_{\min} m^2\ell^2}.$$

Hence, the total regret from cells of length ℓ is at most

$$\begin{aligned}
&(8m^2\ell \max\{1, 1/\lambda_{\max}\} + 5m)|V| \\
&\leq (8m^2\ell \max\{1, 1/\lambda_{\max}\} + 5m) \left(\frac{12 \max\{1, \lambda_{\max}\} \log(T)}{\gamma_{\min} m\ell} + \frac{6\lambda_{\max} \log(T)}{\gamma_{\min} m^2\ell^2} \right) \\
&= \frac{\log(T)}{\gamma_{\min}} \left(96m \max\{\lambda_{\max}, 1/\lambda_{\max}\} + \frac{108 \max\{1, \lambda_{\max}\}}{\ell} + \frac{30\lambda_{\max}}{m\ell^2} \right). \qquad (6)
\end{aligned}$$

Using Eqs. (5) and (6) with $\ell = 2^{-k}$ results in,

$$\begin{aligned}
Reg(T) &\leq T(8m^2 4^{-k} \max\{1, 1/\lambda_{\max}\} + 5m2^{-k}) \\
&\quad + \frac{\log(T)}{\gamma_{\min}} \left(96m \max\{\lambda_{\max}, 1/\lambda_{\max}\} + 108 \max\{1, \lambda_{\max}\} \sum_{i=0}^{k-1} 2^i + \frac{30\lambda_{\max}}{m} \sum_{i=0}^{k-1} 4^i \right) \\
&\leq T(8m^2 4^{-k} \max\{1, 1/\lambda_{\max}\} + 5m2^{-k}) \\
&\quad + \frac{\log(T)}{\gamma_{\min}} \left(96m \max\{\lambda_{\max}, 1/\lambda_{\max}\} + 36 \max\{1, \lambda_{\max}\} 2^k + \frac{10\lambda_{\max}}{m} 4^k \right). \qquad (7)
\end{aligned}$$

for all integer $k \geq 0$. The value of k that minimises regret equalises the leading growth rates of both summands in (7), meaning that $2^k = T^{1/3}$. The claim follows from here.

□

5 Lower Bound on Regret

In this section we give a lower bound on the regret obtained by any algorithm for the filtered Poisson process bandit. The result is given below as Theorem 2, and we see, subject to further minor conditions on the filtering function, that the order of the lower bound on regret matches that of the upper bound on the regret of CIF-UCB up to a logarithmic factor. In this sense, CIF-UCB is therefore asymptotically order optimal (up to the exclusion of logarithmic factors).

Theorem 2. *For the filtered Poisson process bandit problem on* $[0, 1]$ *as described in Section 2 with filtering function* γ *satisfying*

$$\frac{\gamma(a) - \gamma(b)}{b - a} \geq \frac{1}{4}\gamma\left(\frac{a + b}{2}\right) \qquad (8)$$

for any $0 \leq a \leq b \leq 1$*, there exists a valid CIF such that the regret of any algorithm is bounded below as*

$$Reg(T) = \Omega(T^{2/3}).$$

The proof of this lower bound is based on an established analytical technique of relating the regret of an algorithm for a continuum armed bandit problem to that of an algorithm for an associated finite-armed bandit problem. A lower bound on regret for the finite-armed problem is then utilised to lower bound the regret of the continuum armed bandit algorithm.

Here, such an associated finite-armed bandit problem must share the filtering structure of the FPPB to relate regret across the problems, and as such we require a bespoke finite-armed

problem. Therefore, before giving the proof of Theorem 2, we introduce a *filtered Poisson multi-armed bandit* (FPMAB) problem which can be viewed as a discretised version of the FPPB. We derive a lower bound on the regret of any algorithm for the FPMAB, which is a key component of the proof of Theorem 2.

We define the FPMAB problem as follows. The problem is instantiated by K arms with mean parameters $\mu_k \in [0, \lambda_m]$. Each mean parameter may be decomposed as the product of a CIF parameter $\Lambda_k \in [0, \lambda_m]$ and filtering parameter $\gamma_k \in [0, 1]$ - i.e. $\mu_k = \Lambda_k \gamma_k$, $k \in [K]$. The ordered CIF parameters comprise a monotonically increasing sequence, $\Lambda_1 \leq \Lambda_2 \leq \cdots \leq \Lambda_K$, and the ordered filtering parameters comprise a monotonically decreasing sequence, $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K$.

The problem takes place over a series of rounds $t \in [T]$, in each of which the decision-maker selects an arm $a_t \in [K]$ and receives a stochastic reward $R_t = R(a_t)$. In addition, the decision-maker observes *filtered observations*, $\tilde{R}_{k,t}$ for $1 \leq k \leq a_t$. These observations are distributed as

$$\tilde{R}_{k,t} \sim \text{Poisson}(\gamma_{a_t}(\Lambda_k - \Lambda_{k-1})).$$

The reward is defined as the sum of the filtered observations $R_t = \sum_{k=1}^{a_t} \tilde{R}_{k,t}$, and therefore follows a Poisson distribution with parameter μ_a, by the superposition property of the Poisson distribution.

Similarly as to the FPPB, the decision-maker's aim is to minimise regret in T rounds, defined as

$$Reg(T) = \mathbb{E}\Big(\sum_{t=1}^{T} R_t(a^*) - R_t(a_t) \Big),$$

where $a^* \in \text{argmax}_{k \in [K]} \mu_k$ is an optimal arm. We have the following minimax lower bound on the regret of any algorithm for the FPMAB problem.

Theorem 3. *For any number of arms $K \geq 2$, horizon $T \in \mathbb{N}$, a set of filtering parameters $\gamma_1, \ldots, \gamma_K$ satisfiying*

$$\gamma_k \geq \Big(1 + \epsilon\Big)\gamma_{k+1} \tag{9}$$

for $k \in [K-1]$, and some small $\epsilon > 0$ there exist a set of CIF parameters $\Lambda_1, \ldots, \Lambda_K$ and a known constant $C > 0$ such that the regret of any algorithm for the FPMAB problem is at least

$$C\epsilon \left(T - \frac{T}{K} - \frac{T}{2}\sqrt{\frac{3\epsilon^2 T}{K}} \right). \tag{10}$$

This Theorem is similar in spirit to the lower bound on regret for stochastic multi-armed bandits with bounded rewards in Theorem 5.1 of Auer et al. (2002), and its generalisation in Bubeck et al. (2011b). Indeed Theorem 3 has the same order with respect to ϵ and T however there are key differences in the proof of the result. Firstly, Theorem 3 considers filtered Poisson random variables, and therefore parts of the analysis are specific to the KL divergence between two Poisson random variables rather than Bernoulli random variables in the bounded case. Secondly, here we deal with the additional combinatorial feedback of FPMAB problem, and require further analyses to handle the resulting complexities.

In the remainder of this section we prove Theorems 2 and 3.

5.1 Proof of Theorem 2

Proof. Consider the instance of the filtered Poisson process bandit problem referred to as $\mathcal{I}(x^*, \epsilon)$, for $x^* \in [0, 1]$ and $\epsilon > 0$, and specified by the following reward function

$$\nu_{x^*,\epsilon}(x) = \begin{cases} m\epsilon(1 + \epsilon - |x - x^*|), & x : |x - x^*| \leq \epsilon \\ \min(mx, m\epsilon), & \text{othw.} \end{cases} \tag{11}$$

Such a reward function is realised by setting the CIF to

$$\Lambda_{x^*,\epsilon}(x) = \begin{cases} (\gamma(x))^{-1} m\epsilon\big[1 + \epsilon - (x^* - x)\big], & x \in [x^* - \epsilon, x^*) \\ (\gamma(x))^{-1} m\epsilon\big[1 + \epsilon - (x - x^*)\big], & x \in [x^*, x + \epsilon) \\ (\gamma(x))^{-1} \min(mx, m\epsilon), & \text{othw.} \end{cases} \tag{12}$$

To verify that this CIF is increasing, consider the derivative,

$$\frac{d\Lambda_{x^*,\epsilon}(x)}{dx} = \begin{cases} \frac{d(\frac{1}{\gamma})}{dx} m\epsilon\big[1 + \epsilon - x^* + x\big] + m\epsilon(\gamma(x))^{-1}, & x \in [x^* - \epsilon, x^*) \\ \frac{d(\frac{1}{\gamma})}{dx} m\epsilon\big[1 + \epsilon + x^* - x\big] - m\epsilon(\gamma(x))^{-1}, & x \in [x^*, x^* + \epsilon) \\ \frac{d(\frac{1}{\gamma})}{dx} mx + m(\gamma(x))^{-1}, & x \in [0, \epsilon) \\ \frac{d(\frac{1}{\gamma})}{dx} m\epsilon & \text{othw.} \end{cases}$$

We note that $(\gamma(x))^{-1} > 1$ for all $x \in [0, 1]$ since $\gamma : [0, 1] \to [0, 1]$, and that $d(\gamma(x))^{-1}/dx \geq 0$ for all $x \in [0, 1]$ since γ is assumed to be strictly increasing on $[0, 1]$. It follows that for $x \in [x^* - \epsilon, x^*)$,

$$\frac{d\Lambda_{x^*,\epsilon}(x)}{dx} \geq \frac{d(\gamma(x))^{-1}}{dx} m\epsilon\Big[1 + \epsilon - \epsilon\Big] + m\epsilon(\gamma(x))^{-1} = \frac{d(\gamma(x))^{-1}}{dx} m\epsilon + m\epsilon(\gamma(x))^{-1} > 0.$$

For $x \in [x^*, x^* + \epsilon)$, consider

$$\frac{d\Lambda_{x^*,\epsilon}(x)}{dx} \geq \frac{d(\gamma(x))^{-1}}{dx} m\epsilon\Big[1 + \epsilon - \epsilon\Big] - m\epsilon(\gamma(x))^{-1} = m\epsilon\bigg(\frac{d(\gamma(x))^{-1}}{dx} - (\gamma(x))^{-1}\bigg). \tag{13}$$

In the limit as $b - a \to 0$ condition (8) implies that $-\frac{d\gamma(x)}{dx} \geq \gamma(x)$. We have, for a differentiable function f such that $f(x) \neq 0$ that the derivative of $g(x) = 1/f(x)$, that

$$\frac{dg(x)}{dx} = \frac{-\frac{df(x)}{dx}}{(f(x))^2}.$$

Thus,

$$\frac{d(\gamma(x))^{-1}}{dx} = \frac{-\frac{d\gamma(x)}{dx}}{(\gamma(x))^2} \geq \frac{-\gamma(x)}{(\gamma(x))^2} = (\gamma(x))^{-1},$$

and it follows from (13) that $d\Lambda_{x^*,\epsilon}(x)/dx > 0$ for $x \in [x^*, x^* + \epsilon)$. For all other values of $x \in [0, 1]$ it should be obvious that the derivative of the CIF is positive since it comprises a sum of non-negative terms. As such $\Lambda_{x^*,\epsilon}$ satisfies the necessary increasing assumption, and the instance $\mathcal{I}(x^*, \epsilon)$ is a valid instance of the FPPB.

We will lower bound the regret of any algorithm for the problem instance $\mathcal{I}(x^*, \epsilon)$ by relating it to an instance of the filtered Poisson MAB problem.

We fix $K \in \mathbb{N}$ to be defined later and let $\epsilon = (2K)^{-1}$. Further we introduce the function $f_\epsilon : [K] \to [0, 1]$ with

$$f_\epsilon(a) = (2a - 1)\epsilon, \quad a \in [K].$$

This function is used to map between actions in the MAB problem and the CAB problem. We then define an instance $\mathcal{J}(a^*, \epsilon)$ of the K-armed filtered Poisson MAB problem as that with arm means

$$\mu_a = \nu_{x^*,\epsilon}(f_\epsilon(a)), \quad a \in [K],$$

and filtering probabilities

$$\gamma_a = \gamma\left(\frac{2a - 1}{K}\right), \quad a \in [K].$$

It follows that in the problem instance $\mathcal{J}(a^*, \epsilon)$ there is a single optimal arm $a^* \in [K] : x^* \in [\frac{a-1}{K}, \frac{a}{K}]$ with expected reward $\mu_{a^*} = m\epsilon(1 + \epsilon)$ and all other arms, $a \neq a^*$, have expected reward $\mu_a = m\epsilon$.

Let `ALG` be any algorithm for the CAB problem $\mathcal{I}(x^*, \epsilon)$. We will define `ALG'` as an associated algorithm for the MAB problem $\mathcal{J}(a^*, \epsilon)$. These algorithms are related as follows. When `ALG` selects an action $x_t \in [0, 1]$, `ALG'` selects an arm $a_t \equiv a(x_t) \in [K]$ such that

$$x_t \in \left(f_\epsilon(a_t) - \frac{1}{2K}, f_\epsilon(a_t) + \frac{1}{2K}\right).$$

By definition of the FPMAB, `ALG'` will receive reward $R'(a_t) \sim Pois(\mu_{a_t})$ and per-arm observations $\tilde{R}'_{i,t} \sim Pois(\gamma(a_t)(\Lambda_i - \Lambda_{i-1}))$ for $i \leq a_t$. Similarly, `ALG` will receive reward $R(x_t) \sim Pois(\nu_{x^*,\epsilon}(x_t))$ and observe point data in $[0, x_t]$ derived from the filtered Poisson process. We shall also, however, demonstrate that $R(x_t)$ can be shown to have the same distribution as a certain probabilistic function of $\tilde{R}'(a_t)$ and use this representation to relate the regret of `ALG` and `ALG'`.

Define Z to be a Poisson random variable with parameter $m\epsilon(1 + \epsilon)$, and Y to be a Poisson random variable with parameter $m\epsilon$. Then define r_x, a random variable whose distribution depends on $x \in [0, 1]$, as follows,

$$r_x \equiv \begin{cases} Z, & \text{with probability } p_x \\ Y, & \text{othw.} \end{cases} \tag{14}$$

where

$$p_x = \frac{1 - \nu_{x^*,\epsilon}(x)}{1 - \mu_{a(x)}}. \tag{15}$$

It follows that

$$\mathbb{E}(r_x|x) = m\epsilon\left((1 - p_x)\mathbb{E}(Y) + p_x\mathbb{E}(Z)\right)$$

$$
\begin{aligned}
&= m\epsilon\left(1 - \frac{1-\nu_{x^*,\epsilon}(x)}{1-\mu_{a(x)}} + \frac{1-\nu_{x^*,\epsilon}(x)}{1-\mu_{a(x)}}(1+\epsilon)\right) \\
&= m\epsilon\left(1 + \epsilon\ \frac{1-\nu_{x^*,\epsilon}(x)}{1-\mu_{a(x)}}\right) \\
&= \begin{cases} m\epsilon\big(1 + \frac{\epsilon}{1-1-\epsilon}(1-1-\epsilon+m|x-x^*|)\big), & x : |x^*-x| \leq \epsilon \\ m\epsilon, \qquad \text{othw.} \end{cases} \\
&= \mathbb{E}(R(x_t)).
\end{aligned}
$$

We notice that for both $\mathcal{I}(x^*, \epsilon)$ and $\mathcal{J}(a^*, \epsilon)$ the reward of the optimal action is $m\epsilon(1+\epsilon)$. Further we have that $\mathbb{E}(R(x_t)) \leq \mathbb{E}(R'(a(x_t)))$ for all $x_t \in [0,1]$. It therefore follows that the regret of `ALG'` serves as a lower bound on the regret of `ALG`, i.e. we have

$$
\mathbb{E}(Reg_{ALG}(T)) \geq \mathbb{E}(Reg'_{ALG'}(T)).
$$

As `ALG'` is an algorithm for the FPMAB problem, its regret is lower bounded as in Theorem 3, and we therefore have

$$
\mathbb{E}(Reg_{ALG}(T)) \geq C\epsilon\left(T - \frac{T}{K} - \frac{T}{2}\sqrt{\frac{3\epsilon^2 T}{K}}\right),
$$

for a known constant $C > 0$.

We complete the proof of Theorem 2 by optimising our choice of K as a function of T. Substituting $\epsilon = 1/2K$, we have

$$
\mathbb{E}(Reg_{ALG}(T)) \geq \frac{CT}{2K} - \frac{CT}{2K^2} - \frac{CT}{4K}\sqrt{\frac{3T}{4K^3}},
$$

and choosing $K = O(T^{1/3})$ yields the stated result.

□

5.2 Proof of Theorem 3

Proof. Given a set of filtering parameters $\gamma_1, \ldots, \gamma_K$ we construct a problem instance where there is a single "good" arm, $i \in [K]$, with mean reward $\mu_i = 1+\epsilon$, for small $\epsilon \in (0, 1/2]$, and all other arms, $k \neq i$, have mean rewards $\mu_k = 1$. This is achieved by setting the CIF parameters as follows

$$
\Lambda_i^{(i)} = \frac{1+\epsilon}{\gamma_i}, \quad \Lambda_k^{(i)} = \frac{1}{\gamma_k},\ \forall k \neq i.
$$

Here the superscript $\cdot^{(i)}$ denotes that i is the good arm under this choice of parameters, and we notice that the condition of the filtering parameters (9) is required for $\Lambda_1^{(i)}, \ldots \Lambda_K^{(i)}$ to constitute a valid (i.e. increasing) sequence of CIF parameters.

We define three notions of probability and expectation, relevant to the analysis of problem instances of this type. Let $\mathbb{P}_*(\cdot)$ denote probability with respect to the above construction of the FPMAB where the good arm is chosen uniformly at random from $[K]$. Let $\mathbb{P}_i(\cdot)$ be defined

similarly, but denote probability conditioned on the event that $i \in [K]$ is the good arm. Finally let $\mathbb{P}_{equ}$ denote probability with respect to a version where $\mu_k = 1$ for all $k \in [K]$. We let $\mathbb{E}_*(\cdot)$, $\mathbb{E}_i(\cdot)$, and $\mathbb{E}_{equ}(\cdot)$ be respective associated expectation operators.

Let A be the decision-maker's algorithm, let

$$\mathbf{r}_t = (R(a_1), \ldots, R(a_t))$$

denote the sequence of observed rewards in t rounds, and

$$\tilde{\mathbf{r}}_t = \Big((\tilde{R}_{1,1}, \ldots, \tilde{R}_{a_1,1}), \ldots, (\tilde{R}_{1,t}, \ldots, \tilde{R}_{a_t,t})\Big)$$

denote the sequence of filtered observations in t rounds. Any algorithm A may then be thought of a deterministic function from $\{\mathbf{r}_{t-1}, \tilde{\mathbf{r}}_{t-1}\}$ to a_t for all $t \in [T]$. Even an algorithm with randomised action selection can be viewed as deterministic, by treating a given run as a single member of the population of all possible instances of that algorithm.

Further, we define $G_A = \sum_{t=1}^{T} R_t$ to be the reward accumulated by A in T rounds and $G_{max} = \max_{j \in [K]} \sum_{t=1}^{T} R_t(j)$ to be the reward accumulated by playing the best action. The regret of A in T rounds may be expressed as

$$Reg_A(T) = \mathbb{E}\big(G_{max} - G_A\big).$$

Let N_k be the number of times an arm $k \in [K]$ is chosen by A in T rounds. The first step of the proof is to bound the difference in the expectation of N_i when measured using $\mathbb{E}_i$ and $\mathbb{E}_{equ}$, i.e. to bound the difference in the number of times an algorithm with play i between when i is the good arm and when all arms are equally valuable.

Lemma 4. *For any arm i there exists a constant $C(\gamma_{i-1}, \gamma_i, \gamma_{i+1}) > 0$ such that we have*

$$\mathbb{E}_i(N_i) \leq \mathbb{E}_{equ}(N_i) + \frac{T}{2}\sqrt{2\epsilon^2\left(\mathbb{E}_{equ}(N_i)\frac{\gamma_{i-1}}{2(\gamma_{i-1} - \gamma_i)} + F_i\right)}$$

where

$$F_i = C(\gamma_{i-1}, \gamma_i, \gamma_{i+1}) \sum_{k=i+1}^{K} \gamma_k \mathbb{E}_{equ}(N_k), \tag{16}$$

for $\epsilon \leq \frac{\gamma_i}{2\gamma_{i+1}} - \frac{\gamma_i}{2\gamma_{i-1}}$, and $C(\gamma_{i-1}, \gamma_i, \gamma_{i+1})$ is a known positive constant.

By construction of the CIF paramters $\Lambda_1^{(i)}, \ldots, \Lambda_K^{(i)}$ we have that for any $t \in [T]$, $\mathbb{E}(R_t) = 1 + \epsilon\mathbb{P}_i(a_t = i)$. It follows that the expected reward of algorithm A, G_A satisfies $\mathbb{E}_i(G_A) = T + \epsilon\mathbb{E}_i(N_i)$. The expectation in the regret measure is taken with respect to $\mathbb{P}_*$, rather than any $\mathbb{P}_i$, as such $\mathbb{E}_*(G_A)$ is the quantity of interest. We recall that under $\mathbb{P}_*$ the "good" arm is chosen uniformly at random, and thus, it follows that

$$\mathbb{E}_*(G_A) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_k(G_a) \leq T + \frac{1}{K}\sum_{k=1}^{K}\epsilon\mathbb{E}_k(N_k)$$

$$\leq T + \frac{\epsilon}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{equ}(N_k) + \frac{T}{2}\sqrt{2\epsilon^2\left(\mathbb{E}_{equ}(N_k)\frac{\gamma_{k-1}}{2(\gamma_{k-1}-\gamma_k)} + F_k\right)}\right)$$
$$= T + \frac{\epsilon T}{K} + \frac{\epsilon T}{2K}\sum_{k=1}^{K}\sqrt{2\epsilon^2\left(\mathbb{E}_{equ}(N_k)\frac{\gamma_{k-1}}{2(\gamma_{k-1}-\gamma_k)} + F_k\right)}, \tag{17}$$

where the second inequality uses Lemma 4.

Considering the final term of (17), we have by Cauchy-Schwarz,

$$\sum_{k=1}^{K}\sqrt{2\epsilon^2\left(\mathbb{E}_{equ}(N_k)\frac{\gamma_{k-1}}{2(\gamma_{k-1}-\gamma_k)} + F_k\right)} \leq \sqrt{K\sum_{k=1}^{K}2\epsilon^2\left(\mathbb{E}_{equ}(N_k)\frac{\gamma_{k-1}}{2(\gamma_{k-1}-\gamma_k)} + F_k\right)}$$
$$\leq \sqrt{\epsilon^2 KT + 2\epsilon^2 K\sum_{k=1}^{K}C(\gamma_{k-1},\gamma_k,\gamma_{k+1})\sum_{j=k}^{K}\gamma_j\mathbb{E}_{equ}(N_j)}$$
$$\leq \sqrt{3\epsilon^2 KT\max_{k\in[K]}C(\gamma_{k-1},\gamma_k,\gamma_{k+1})}$$

Thus

$$\mathbb{E}_*(G_A) \leq T + \frac{\epsilon T}{K} + \frac{\epsilon T}{2}\sqrt{\frac{3\epsilon^2 T\max_{k\in[K]}C(\gamma_{k-1},\gamma_k,\gamma_{k+1})}{K}},$$

and the regret is bounded as

$$\mathbb{E}_*|G_{max} - G_A| \geq (1+\epsilon)T - T - \frac{\epsilon T}{K} - \frac{\epsilon T}{2}\sqrt{\frac{3\epsilon^2 T\max_{k\in[K]}C(\gamma_{k-1},\gamma_k,\gamma_{k+1})}{K}}$$
$$= \epsilon T - \frac{\epsilon T}{K} - \frac{\epsilon T}{2}\sqrt{\frac{3\epsilon^2 T\max_{k\in[K]}C(\gamma_{k-1},\gamma_k,\gamma_{k+1})}{K}}.$$

□

5.3 Proof of Lemma 4

We first introduce some further notation used in the proof. Define for any distributions $\mathbb{P}$ and $\mathbb{Q}$ over vector sequences $\tilde{\mathbf{r}} \in \mathbb{N}^{K\times T}$, the variational distance as

$$||\mathbb{P} - \mathbb{Q}||_1 \equiv \sum_{\mathbf{r}\in\mathbb{N}^{K\times T}}|\mathbb{P}(\tilde{\mathbf{r}}) - \mathbb{Q}(\tilde{\mathbf{r}})|,$$

and the KL divergence as

$$KL(\mathbb{P}\;||\;\mathbb{Q}) \equiv \sum_{\mathbf{r}\in\mathbb{N}^{K\times T}}\mathbb{P}(\mathbf{r})\log\left(\frac{\mathbb{P}(\mathbf{r})}{\mathbb{Q}(\mathbf{r})}\right).$$

By Pinsker's inequality, we have the following relationship between these distances

$$||\mathbb{P} - \mathbb{Q}||_1 \leq \sqrt{2KL(\mathbb{Q}\;||\;\mathbb{P})}. \tag{18}$$

Finally, the KL divergence between two Poisson distributions with parameters λ and ν is given as,

$$KL(\lambda||\nu) \equiv \lambda \log\left(\frac{\lambda}{\nu}\right) + \nu - \lambda.$$

Proof. For any function $f : \mathbb{N}^{K\times T} \to [0, M]$, with $M > 0$ constant, we have,

$$\begin{aligned}
\mathbb{E}_i(f(\tilde{\mathbf{r}})) - \mathbb{E}_{equ}(f(\tilde{\mathbf{r}})) &= \sum_{\tilde{\mathbf{r}}\in\mathbb{N}_+^{K\times T}} f(\tilde{\mathbf{r}})\big(\mathbb{P}_i(\tilde{\mathbf{r}}) - \mathbb{P}_{equ}(\tilde{\mathbf{r}})\big) \\
&\leq \sum_{\tilde{\mathbf{r}}:\mathbb{P}_i(\tilde{\mathbf{r}})\geq\mathbb{P}_{equ}(\tilde{\mathbf{r}})} f(\tilde{\mathbf{r}})\big(\mathbb{P}_i(\tilde{\mathbf{r}}) - \mathbb{P}_{equ}(\tilde{\mathbf{r}})\big) \\
&\leq \frac{M}{2}||\mathbb{P}_i - \mathbb{P}_{equ}||_1 \\
&\leq \frac{M}{2}\sqrt{2KL(\mathbb{P}_{equ}||\mathbb{P}_i)}, \qquad (19)
\end{aligned}$$

where the final inequality follows from (18). Considering the KL divergence term in isolation, we have, by Theorem 2.5.3 of Cover and Thomas (2012)

$$\begin{aligned}
KL(\mathbb{P}_{equ} \,||\, \mathbb{P}_i) &= \sum_{t=1}^{T} KL\big(\mathbb{P}_{equ}(\tilde{\mathbf{r}}_t \mid \tilde{\mathbf{r}}_{1:t-1}) \,\big|\big|\, \mathbb{P}_i(\tilde{\mathbf{r}}_t \mid \tilde{\mathbf{r}}_{1:t-1})\big) \\
&= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{P}_{equ}\big(a_t = k\big) KL\big(\mathbb{P}_{equ}(\tilde{\mathbf{r}}_t \mid a_t = k) \,\big|\big|\, \mathbb{P}_i(\tilde{\mathbf{r}}_t \mid a_t = k)\big) \\
&= \sum_{t=1}^{T}\sum_{k=i}^{K} \mathbb{P}_{equ}\big(a_t = k\big) KL\big(\mathbb{P}_{equ}(\tilde{\mathbf{r}}_t \mid a_t = k) \,\big|\big|\, \mathbb{P}_i(\tilde{\mathbf{r}}_t \mid a_t = k)\big) \\
&= \sum_{t=1}^{T}\sum_{k=i}^{K} \mathbb{P}_{equ}\big(a_t = k\big) \sum_{j=1}^{k} KL\Big(\gamma_k(\Lambda_j^{equ} - \Lambda_{j-1}^{equ}), \gamma_k(\Lambda_j^{(i)} - \Lambda_{j-1}^{(i)})\Big) \\
&= \sum_{t=1}^{T}\sum_{k=i}^{K} \mathbb{P}_{equ}\big(a_t = k\big) \sum_{j=1}^{k} KL\Big(\gamma_k(\frac{1}{\gamma_j} - \frac{1}{\gamma_{j-1}}), \gamma_k(\Lambda_j^{(i)} - \Lambda_{j-1}^{(i)})\Big).
\end{aligned}$$

Here the parameters Λ_k^{equ}, $k \in [K]$ refer to the choice of CIF parameters which yields $\mu_k = 1$ for all $k \in [K]$. The final equality follows from the observation that if $a_t < i$ then the distribution of the filtered observations is identical under $\mathbb{P}_{equ}$ and $\mathbb{P}_i$. Decomposing on the sum over k, with the observation that for $j > i + 1$ the CIF parameters under the "single good arm" and "all arms equal" constructions will also match, meaning $KL(\gamma_k(\Lambda_j^{equ} - \Lambda_{j-1}^{equ}), \gamma_k(\Lambda_k^{(i)} - \Lambda_{j-1}^{(i)})) = 0$, for any $j > i + 1$ we have

$$\begin{aligned}
&KL(\mathbb{P}_{equ} \,||\, \mathbb{P}_i) \\
&= \sum_{t=1}^{T} \mathbb{P}_{equ}(a_t = i) KL\Big(\gamma_i(\frac{1}{\gamma_i} - \frac{1}{\gamma_{i-1}}), \gamma_i(\frac{1+\epsilon}{\gamma_i} - \frac{1}{\gamma_{i-1}})\Big) \\
&\quad + \sum_{t=1}^{T}\sum_{k=i+1}^{K} \mathbb{P}_{equ}\big(a_t = k\big) \sum_{j\in\{i,i+1\}} KL\Big(\gamma_k(\frac{1}{\gamma_j} - \frac{1}{\gamma_{j-1}}), \gamma_k(\Lambda_j^{(i)} - \Lambda_{j-1}^{(i)})\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{equ}(N_i)\left((1-\frac{\gamma_i}{\gamma_{i-1}})\log\left(\frac{1-\frac{\gamma_i}{\gamma_{i-1}}}{1+\epsilon-\frac{\gamma_i}{\gamma_{i-1}}}\right)+\epsilon\right) \\
&\quad + \sum_{k=i+1}^{K}\mathbb{E}_{equ}(N_k)\Bigg[\left((\frac{\gamma_k}{\gamma_i}-\frac{\gamma_k}{\gamma_{i-1}})\log\left(\frac{\frac{1}{\gamma_i}-\frac{1}{\gamma_{i-1}}}{\frac{1+\epsilon}{\gamma_i}-\frac{1}{\gamma_{i-1}}}\right)+\epsilon\frac{\gamma_k}{\gamma_i}\right) \\
&\qquad\qquad + \left((\frac{\gamma_k}{\gamma_{i+1}}-\frac{\gamma_k}{\gamma_i})\log\left(\frac{\frac{1}{\gamma_{i+1}}-\frac{1}{\gamma_i}}{\frac{1}{\gamma_{i+1}}-\frac{1+\epsilon}{\gamma_i}}\right)-\epsilon\frac{\gamma_k}{\gamma_i}\right)\Bigg] \\
&\leq \mathbb{E}_{equ}(N_i)\frac{\gamma_{i-1}}{2(\gamma_{i-1}-\gamma_i)}\epsilon^2 \\
&\quad + \sum_{k=i+1}^{K}\mathbb{E}_{equ}(N_k)\frac{\gamma_k}{\gamma_i}\left[\frac{\gamma_{i-1}-\gamma_i}{\gamma_{i-1}}\log\left(\frac{1}{1+\frac{\gamma_{i-1}}{\gamma_{i-1}-\gamma_i}\epsilon}\right)+\frac{\gamma_i-\gamma_{i+1}}{\gamma_{i+1}}\log\left(\frac{1}{1-\frac{\gamma_{i+1}}{\gamma_i-\gamma_{i+1}}\epsilon}\right)\right], \quad (20)
\end{aligned}
$$

for $\epsilon \leq \frac{\gamma_i-\gamma_{i+1}}{\gamma_{i+1}}$. The inequality uses the identity

$$
2ax + 2a^2\log\left(\frac{a}{a+x}\right) \leq x^2, \;\; x>0, \;\; a<1.
$$

It remains to bound the summation in (20) with an $o(\epsilon^2)$ term. For general $a \in [0,1]$, $b \in [0,1]$, and $0 \leq x \leq b$, consider the function

$$
g(x) = a\log\left(\frac{1}{1+\frac{x}{a}}\right) + b\log\left(\frac{1}{1+\frac{x}{b}}\right).
$$

We have its derivative

$$
\frac{dg(x)}{dx} = \frac{a}{a-x} + \frac{b}{x-b},
$$

and thus for some $C>1$ we have the following linear bound on the derivative

$$
dg/dx \leq 2Cx, \quad \forall\, x \in \left[0, \frac{a+b}{2} + \sqrt{\frac{(a+b)^2}{4} - \frac{ab+(a-b)}{4C}}\right]. \quad (21)
$$

Solutions to $g(x) = Cx^2$ are not available in closed-form, but since $g(0)=0$, and $dg/dx|_{x=0} = 0$ we have as a minimum that $g(x) \leq Cx^2$ for x as in (21). Choosing $C = \frac{ab+a-b}{(a+b)^2}$ gives $g(x) \leq Cx^2$ for $x \in [0, \frac{a+b}{2}]$.

It therefore follows that

$$
\begin{aligned}
&\frac{\gamma_{i-1}-\gamma_i}{\gamma_{i-1}}\log\left(\frac{1}{1+\frac{\gamma_{i-1}}{\gamma_{i-1}-\gamma_i}\epsilon}\right)+\frac{\gamma_i-\gamma_{i+1}}{\gamma_{i+1}}\log\left(\frac{1}{1-\frac{\gamma_{i+1}}{\gamma_i-\gamma_{i+1}}\epsilon}\right) \\
&\qquad \leq \left(\frac{\gamma_i(\gamma_{i-1}-2\gamma_i+\gamma_{i+1})}{\frac{\gamma_{i+1}}{\gamma_{i-1}}(\gamma_{i-1}-\gamma_i)^2+2(\gamma_{i-1}-\gamma_i)(\gamma_i-\gamma_{i+1})+\frac{\gamma_{i-1}}{\gamma_{i+1}}(\gamma_i-\gamma_{i+1})^2}\right)\epsilon^2, \quad (22)
\end{aligned}
$$

for all $x \in \left[0, \frac{\gamma_i}{2\gamma_{i+1}} - \frac{\gamma_i}{2\gamma_{i-1}}\right]$.

Combining (20) and (22) we therefore have that the KL divergence from $\mathbb{P}_{equ}$ to $\mathbb{P}_i$ may be

bounded as follows,

$$KL(\mathbb{P}_{equ} \mid\mid \mathbb{P}_i) \leq \epsilon^2 \left(\frac{\gamma_{i-1}}{2(\gamma_{i-1} - \gamma_i)} \mathbb{E}_{equ}(N_i) + C(\gamma_{i-1}, \gamma_i, \gamma_{i+1}) \sum_{k=i+1}^{K} \gamma_k \mathbb{E}_{equ}(N_k), \right) \tag{23}$$

for $\epsilon \leq \frac{\gamma_i}{2\gamma_{i+1}} - \frac{\gamma_i}{2\gamma_{i-1}}$, where $C(\gamma_{i-1}, \gamma_i, \gamma_{i+1})$ is a known positive constant. Finally, as $N_i : \mathbb{N}^{K \times T} \to [0, T]$, we have the stated result by the combination of (19), and (23).

□

6 Experiments

In this section we illustrate the performance of CIF-UCB via numerical examples. We work with a linear intensity function $\lambda(x) = 20 - 20x$ and exponential filtering probability $\gamma(x) = \exp(-x)$, both for $x \in [0, 1]$. The plot of $\Lambda(x)\gamma(x)$ is shown in Figure 2, with $x^* = 0.586$ and $\Lambda(x^*)\gamma(x^*) = 4.61$ (found numerically). In the experiment, we set the Lipschitz constant $m = 20$, which equals $\max_{0 \leq x \leq 1}(\Lambda(x)\gamma(x))'$ (since $\Lambda(x)\gamma(x)$ is concave), and $\lambda_{\max} = 20$.

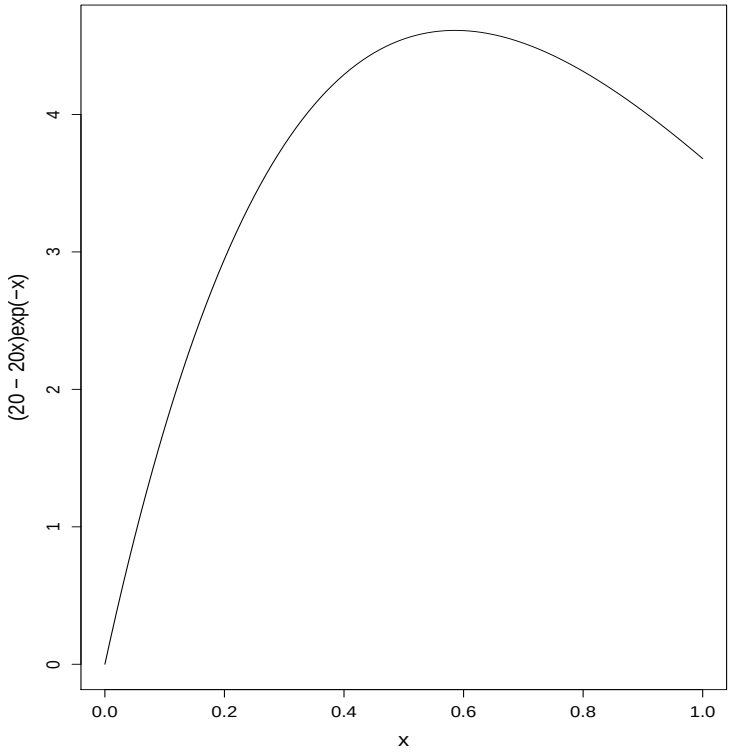

Figure 2: Plot of $\Lambda(x)\gamma(x)$.

We ran 100 independent sample paths over a time horizon of $T = 50000$, and computed the average cumulative regret over the 100 sample paths. The resulting average cumulative regret is shown in Figure 3, along with the upper regret bound, as determined in Theorem 1.

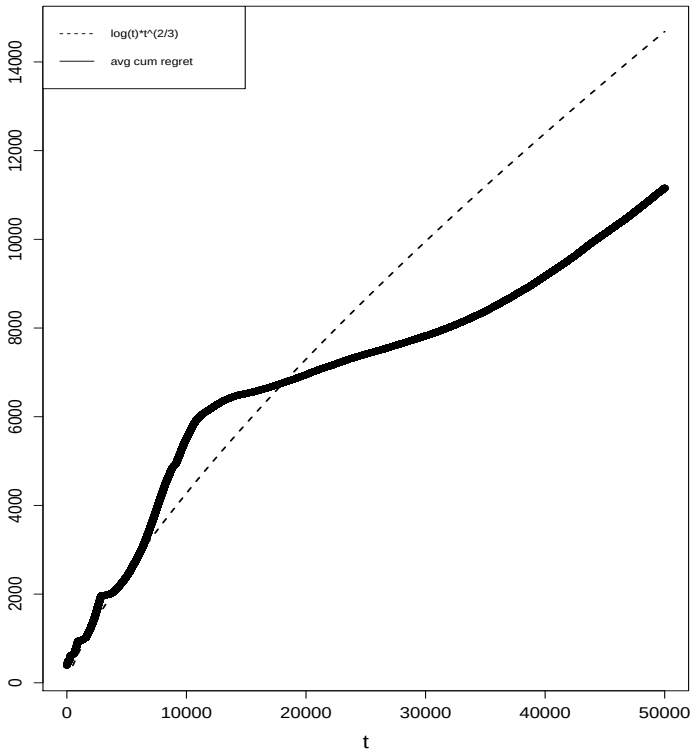

Figure 3: Plot of average cumulative regret.

Several observations are in order. First, the dotted curve in Figure 3 doesn't include the constant terms (equal to 360 in this case) nor the sub $\log(t)t^{2/3}$ terms that come up in the regret upper bound derivation (cf. Eq. (7)). Still, we note that the regret growth is plausibly of order $\tilde{O}(t^{2/3})$.

The second observation concerns the shape of the average cumulative regret. Note that the cumulative regret appears to be piece-wise convex increasing, such that the regret of each extra convex piece grows at a slower rate; this observation is even more noticeable on individual sample paths (not shown). This growth pattern is due to the splitting condition of CIF-UCB, whereby the algorithm initially samples the best of the two segments that result from a split, and explores other (typically worse) segments as t gets larger. As t grows, the algorithm exploits more often, and thus each convex piece grows slower.

The final observation is about the splitting pattern. We include in Table 6 the data frame for the final round of a sample path in the R implementation, which includes the two endpoints (x and y), the effective number of samples of each final segment $\sum_{i=1}^{|V_T(x,y)|} \gamma(b_{\tau_i})$, the index $\mathcal{I}_T(x,y)$, and the CIF estimator $\bar{\Lambda}_T(y)$ in the rightmost column. The finer spatial grid around x^* is appreciable, suggesting that the algorithm gravitates towards the segment that contains the optimal solution x^*. Note also that the estimates of $\Lambda(x) = 20x - 10x^2$ are very precise (the largest relative error is .62% for y large, since the segments close to 1 have the fewest number of effective samples $\sum_{i=1}^{|V_T(x,y)|} \gamma(b_{\tau_i})$). The index values are similar across the final segments, as is

typical with UCB algorithms, and the effective number of samples drops off significantly to the right of x^*. On the other hand, the effective number of samples to the left of x^* is large, since the algorithm needs to cover that space to reach (and exploit) the neighborhood around x^*.

x	y	$\sum_{i=1}^{\lvert V_T(x,y)\rvert} \gamma(b_{\tau_i})$	$\mathcal{I}_T(x,y)$	$\bar{\Lambda}_T(y)$
0.0000000	0.1250000	24118.996	4.740275	2.348149
0.1250000	0.1875000	24118.996	4.225897	3.398939
0.1875000	0.2500000	24118.996	4.802038	4.370248
0.2500000	0.2812500	24117.438	4.413122	4.827627
0.2812500	0.3125000	24117.438	4.615670	5.263785
0.3125000	0.3437500	24116.707	4.794795	5.689541
0.3437500	0.3750000	24116.707	4.943961	6.093245
0.3750000	0.3906250	24115.332	4.692382	6.282144
0.3906250	0.4062500	24115.332	4.747511	6.466840
0.4062500	0.4218750	24114.666	4.797825	6.648402
0.4218750	0.4375000	24114.666	4.843334	6.826593
0.4375000	0.4531250	24113.375	4.882826	6.999228
0.4531250	0.4687500	24113.375	4.916710	7.166604
0.4687500	0.4843750	24112.749	4.946188	7.330313
0.4843750	0.4921875	24112.749	4.803775	7.411888
0.4921875	0.5000000	24112.749	4.814192	7.488694
0.5000000	0.5078125	23996.902	4.825221	7.566727
0.5078125	0.5156250	23996.902	4.834771	7.643570
0.5156250	0.5234375	23996.304	4.845463	7.723064
0.5234375	0.5312500	23996.304	4.852431	7.796992
0.5312500	0.5390625	23995.129	4.858238	7.869597
0.5390625	0.5468750	23995.129	4.861613	7.938653
0.5468750	0.5546875	23994.550	4.865409	8.009027
0.5546875	0.5625000	23994.550	4.868062	8.078001
0.5625000	0.5703125	23875.465	4.869991	8.145726
0.5703125	0.5781250	23875.465	4.870570	8.212154
0.5781250	0.5859375	23726.253	4.870049	8.276444
0.5859375	0.5937500	23726.253	4.869350	8.341604
0.5937500	0.6015625	23314.273	4.869802	8.407425
0.6015625	0.6093750	23314.273	4.867215	8.470133
0.6093750	0.6171875	21772.366	4.866196	8.528425
0.6171875	0.6250000	21772.366	4.861871	8.588823
0.6250000	0.6328125	20964.657	4.860349	8.650464
0.6328125	0.6406250	20964.657	4.854130	8.708132
0.6406250	0.6484375	18021.041	4.851132	8.753434
0.6484375	0.6562500	18021.041	4.843125	8.808425
0.6562500	0.6640625	16849.088	4.842293	8.868670
0.6640625	0.6718750	16849.088	4.833623	8.923094
0.6718750	0.6796875	14015.447	4.831297	8.963610
0.6796875	0.6875000	14015.447	4.820711	9.014911
0.6875000	0.6953125	11653.144	4.823237	9.062533
0.6953125	0.7031250	11653.144	4.811218	9.111618
0.7031250	0.7187500	7729.984	4.954881	9.151894
0.7187500	0.7343750	6535.462	4.945111	9.240815
0.7343750	0.7500000	6535.462	4.910583	9.319769
0.7500000	0.7656250	6072.071	4.885203	9.399100
0.7656250	0.7812500	6072.071	4.848842	9.474527
0.7812500	0.7968750	5608.286	4.820870	9.546411
0.7968750	0.8125000	5608.286	4.778192	9.607749
0.8125000	0.8281250	4692.391	4.770177	9.693565
0.8281250	0.8437500	4692.391	4.723798	9.746417
0.8437500	0.8593750	3776.720	4.714856	9.810630
0.8593750	0.8750000	3776.720	4.664373	9.853260
0.8750000	0.9062500	2290.606	4.954441	9.899998
0.9062500	0.9375000	1757.278	4.886901	9.906228
0.9375000	0.9687500	1236.443	4.740351	9.937378
0.9687500	1.0000000	1236.443	4.746599	9.954363

Table 1: Summary of main parameters after a sample path

To test the sensitivity of the algorithm to multiple local maximums, we ran a second experiment with parameters identical to those of the first experiment, except for the filtering

probability $\gamma(\cdot)$, which now is set to be piece-wise linearly decreasing,

$$\gamma(x) = \begin{cases} 1, \text{ for } x \in [0, 0.25) \\ 1.5 - 2x \text{ for } x \in [0.25, 0.5) \\ 0.5, \text{ for } x \in [0.5, 0.8) \\ 1.3 - x \text{ for } x \in [0.8, 1]. \end{cases}$$

This filtering probability leads to a $\Lambda(x)\gamma(x)$ objective as in Figure 4, with $x^* = 0.8$ and $\Lambda(x^*)\gamma(x^*) = 4.8$.

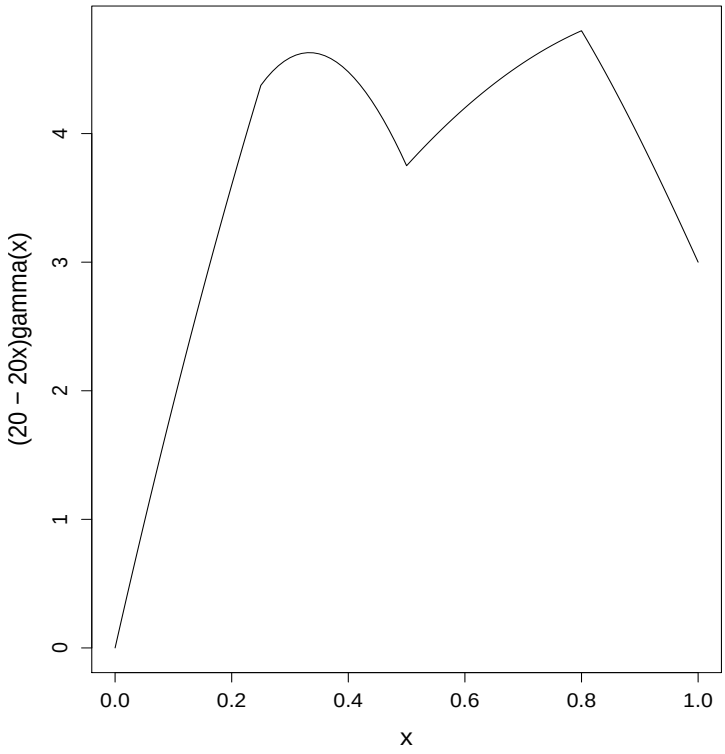

Figure 4: Plot of $\Lambda(x)\gamma(x)$.

We tested CIF-UCB over 100 independent sample paths, with a time horizon $T = 50000$. This resulted in an average cumulative regret as shown in Figure 5.

Two main observations can be drawn. First, the $\tilde{O}(T^{2/3})$ upper bound of Theorem 1 holds over $t \in \{1, \ldots, T\}$. Second, the average cumulative regret is about 10% larger than in the first experiment for $t = T$. This can be ascribed to the fact that the optimal value of the objective function is 4.8 versus 4.61 in the first experiment, and to the extra exploration induced by the local maximum at $x = .33$.

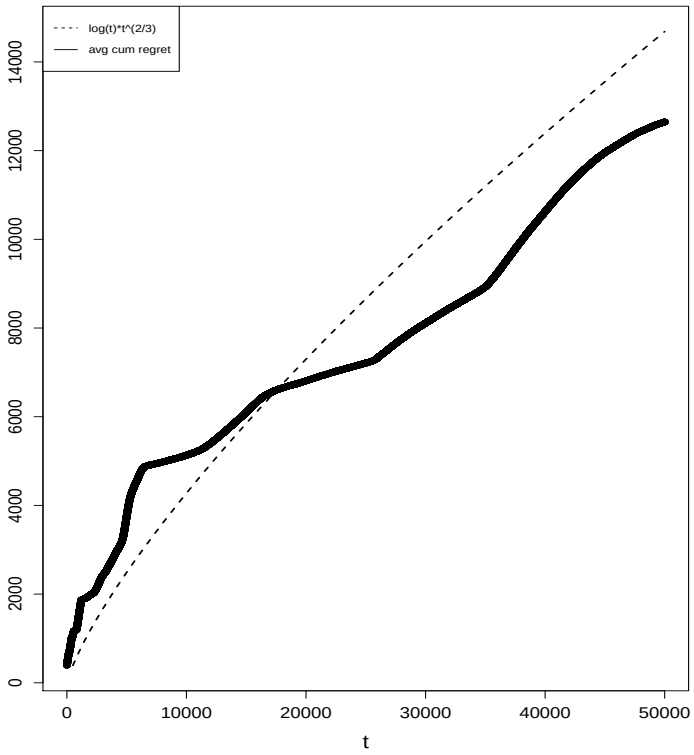

Figure 5: Plot of average cumulative regret.

7 Discussion

This work considers a sequential variant of the problem faced by a decision-maker who attempts to maximise the detection of events generated by a filtered non-homogeneous Poisson process, where the filtering probability depends on the segment selected by the decision-maker, and the Poisson cumulative intensity function is unknown. The independent increment property of the Poisson process makes the analysis tractable, enabling the use of the machinery developed for the continuum bandit problem. The problem of efficient exploration/exploitation of a filtered Poisson process on a continuum arises naturally in settings where observations are made by searchers (representing cameras, sensors, robotic and human searchers, etc.), and the events that generate observations tend to disappear (or renege, in a queueing context), before an observation can be made, as the interval of search increases. Besides extending the state-of-the-art to such settings, the main contributions are an algorithm for a filtered Poisson process on a continuum, and regret bounds that are optimal up to a logarithmic factor.

Acknowledgements JAG was supported by EPSRC grant EP/L015692/1 (STOR-i Centre for Doctoral Training). RS was supported by ONR grant N0001420WX00860.

References

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011a). X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695.

Bubeck, S., Stoltz, G., and Yu, J. Y. (2011b). Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Grant, J. A., Boukouvalas, A., Griffiths, R.-R., Leslie, D. S., Vakili, S., and De Cote, E. M. (2019). Adaptive sensor placement for continuous spaces. *International Conference on Machine Learning*.

Grant, J. A., Leslie, D. S., Glazebrook, K., Szechtman, R., and Letchford, A. N. (2020). Adaptive policies for perimeter surveillance problems. *European Journal of Operational Research*, 283:265–278.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690.

Kleinberg, R. D. (2005). Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704.

Lu, S., Wang, G., Hu, Y., and Zhang, L. (2019). Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning*, pages 4154–4163.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*. Omnipress.